# BIOS 545: Statistical Analysis

Dane Van Domelen

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta, GA

April 9, 2018

# Statistical analysis in R

- Analyzing data is what (most) statisticians do.
- General procedure:
    1. Load data into R.
    2. Clean data (boring am I right??)
    3. **Try to answer a research question.**

# Format of lecture

- **Basic idea:** Go through various research questions on a particular dataset to illustrate data analysis tools in R.
- **Dataset:** NHANES physical activity.
- **General procedure for each research question:**

    (1) Visualize data (review!)

    (2) Estimate parameter of interest

    (3) Perform hypothesis test

# NHANES dataset

- National Health And Nutrition Examination Survey
  - Cross-sectional study in the US.
  - $n \approx 10,000$ in each 2-year cycle.
  - Demographics, questionnaires, lab tests, etc.
  - Publicly available!
    `https://www.cdc.gov/nchs/nhanes/`

# Putting dataset together (FYI)

```r
# Install/load packages
install_github("vandomed/nhanesaccel")
install_github("vandomed/nhanesdata")
library("nhanesaccel")
library("nhanesdata")

# Process NHANES 2003-2006 data
nhanes.pa <- process_nhanes(waves = 1, valid_wk_days = 5, valid_we_days = 2,
                            weekday_weekend = TRUE, brevity = 2)

# Subset participants with usable data, and only keep certain variables
nhanes.pa <- nhanes.pa %>%
  filter(include == 1) %>%
  select(seqn, cpm, sed_min, guideline_min, wk_cpm, we_cpm)

# Merge in demographics and body measurements datasets
data(demo_c)
data(bmx_c)
names(demo_c) <- tolower(names(demo_c))
names(bmx_c) <- tolower(names(bmx_c))
nhanes <- nhanes.pa %>% inner_join(demo_c) %>% inner_join(bmx_c)

# Subset variables of interest and create some factors
nhanes <- nhanes %>%
  select(seqn, riagendr, ridageyr, ridreth2, dmdeduc2, indfmpir, bmxbmi,
         bmxwaist, cpm, sed_min, guideline_min, wk_cpm, we_cpm) %>%
  rename(sex = riagendr, age = ridageyr, eth = ridreth2, educ = dmdeduc2,
         pir = indfmpir, bmi = bmxbmi, waist = bmxwaist) %>%
  mutate(sex = factor(sex, levels = c(1, 2), labels = c("Male", "Female")))
```

# Initial look at dataset

```
# Download dataset from website
load(url(
  "https://github.com/vandomed/vandomed.github.io/raw/master/nhanes.rda"))
```

```
# Look at data structure
dim(nhanes)

[1] 7176   13
```

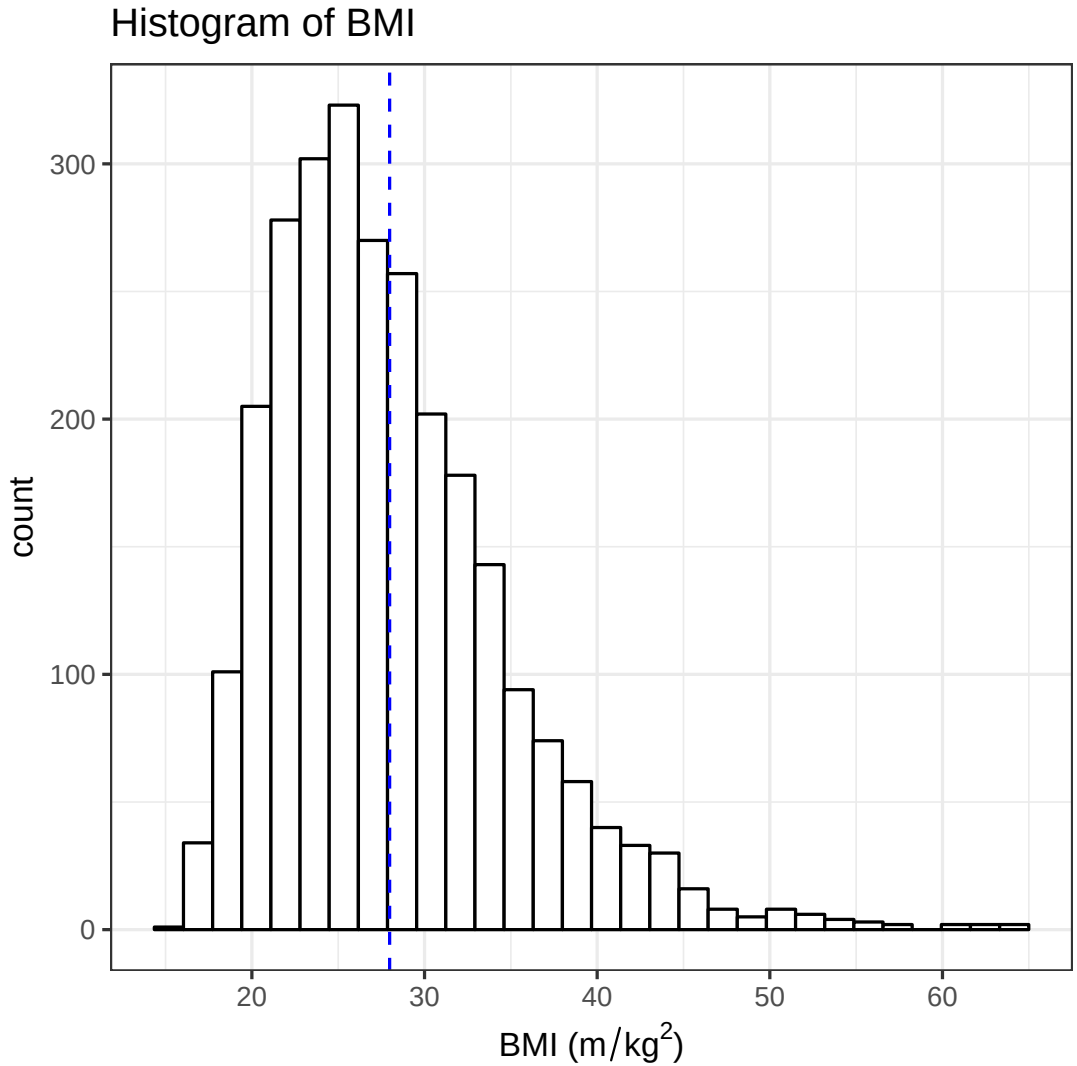
```
head(nhanes, 3)

   seqn    sex age eth educ  pir   bmi waist    cpm sed_min guideline_min
1 21005   Male  19   2   NA 2.44 50.85 135.9 609.54  458.67        19.333
2 21006 Female  16   2   NA 2.47 20.78  73.6 145.46  591.50         2.750
3 21007 Female  14   1   NA 1.60 18.43  69.5 402.28  482.71         0.000
  wk_cpm we_cpm
1 649.29 530.04
2 171.18 119.74
3 366.36 492.09
```

**Research question:** What is the mean BMI of American adults age 18-50?

# Visualization

```r
# Get subset of participants age 18-50
nhanes.adults <- subset(nhanes, age >= 18 & age <= 50)

# Create histogram
ggplot(nhanes.adults, aes(x = bmi)) +
  geom_histogram(col = "black", fill = "white") +
  labs(title = "Histogram of BMI",
       x = expression(paste("BMI (", m/kg^2, ")"))) +
  geom_vline(aes(xintercept = mean(bmi, na.rm = TRUE)),
             color = "blue", linetype = 2) +
  theme_bw()
```

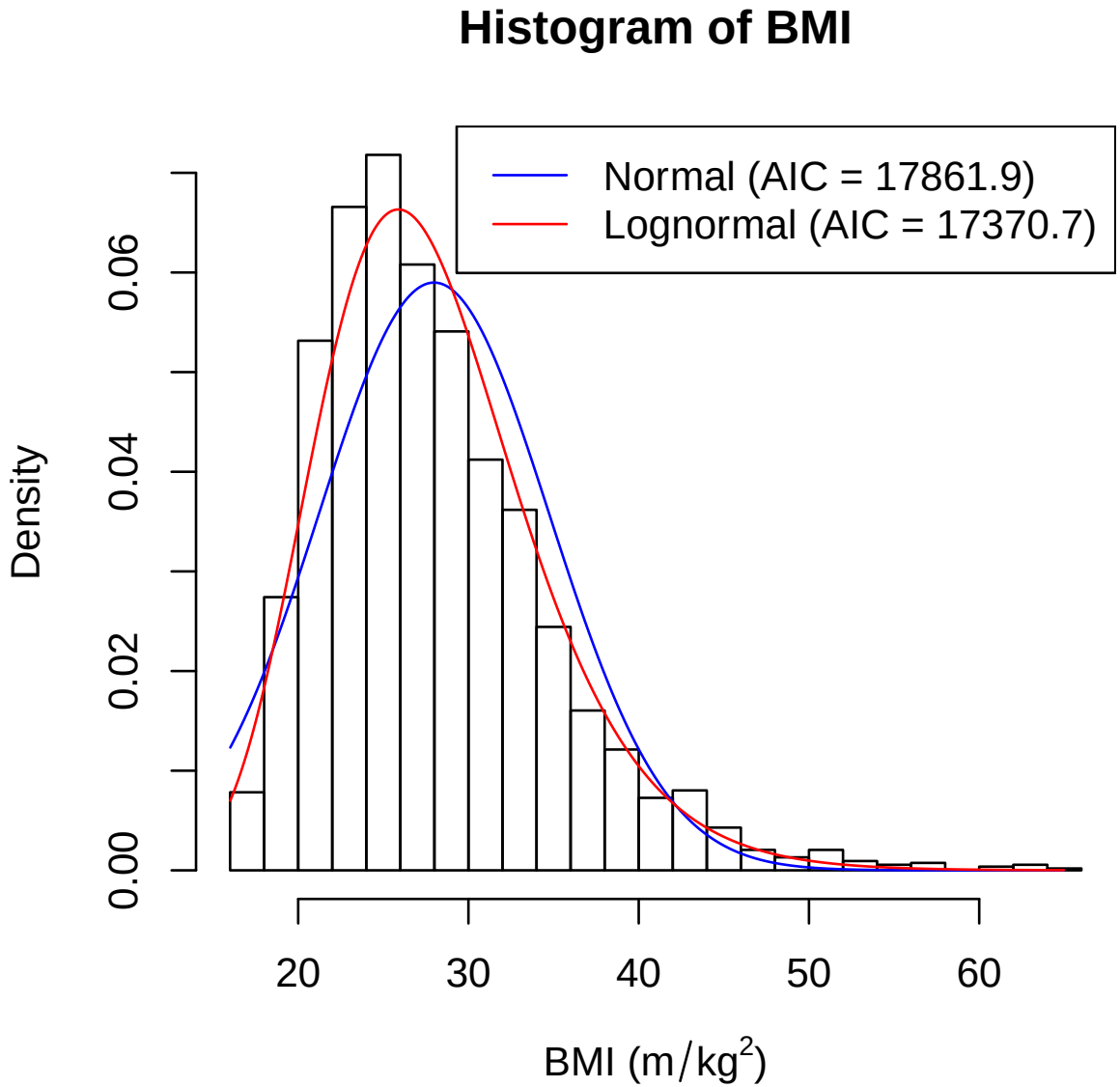
# Visualization



# Normal or lognormal?

```
# Create histogram with densities overlaid
library("dvmisc")
histo(nhanes.adults$bmi,
      breaks = 20,
      main = "Histogram of BMI",
      xlab =  expression(paste("BMI (", m/kg^2, ")")),
      dis = c("norm", "lnorm"),
      colors = c("blue", "red"),
      lty = c(1, 1),
      legend_form = 2, aic_decimals = 1)
```

# Normal or lognormal?



# Parameter estimation

- Statistical setup:
  - Let $X =$ BMI.
  - Assume $X_1, ..., X_n \overset{iid}{\sim} (\mu, \sigma^2)$
- Estimators:

  $\hat{\mu} = \bar{X}$

  95% CI for $\mu$ : $\bar{X} \pm \frac{t_{(.975, n-1)} s}{\sqrt{n}}$

# Parameter estimation

```r
# Calculate sample mean
(x.bar <- mean(nhanes.adults$bmi, na.rm = T))
```

```
[1] 27.979
```

```r
# Calculate 95% CI manually
s <- sd(nhanes.adults$bmi, na.rm = T)
n <- sum(! is.na(nhanes.adults$bmi))
t <- qt(p = 0.975, df = n - 1)
c(x.bar - t * s / sqrt(n), x.bar + t * s / sqrt(n))
```

```
[1] 27.723 28.236
```

# Parameter estimation

```
# Better to use built-in R function!
t.test(nhanes.adults$bmi)
```

```
One Sample t-test

data:  nhanes.adults$bmi
t = 214, df = 2680, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 27.723 28.236
sample estimates:
mean of x
   27.979
```

# Hypothesis test

Suppose we want to test:

$$H_0 : \mu = 25$$

$$H_A : \mu \neq 25$$

# Hypothesis testing

```
(ttest.fit <- t.test(nhanes.adults$bmi, mu = 25))
```

```
One Sample t-test

data:  nhanes.adults$bmi
t = 22.8, df = 2680, p-value <2e-16
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 27.723 28.236
sample estimates:
mean of x
   27.979
```

# Hypothesis testing

```
names(ttest.fit)
```

```
[1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
[6] "null.value"  "alternative" "method"      "data.name"
```

```
ttest.fit$estimate
```

```
mean of x
   27.979
```

```
ttest.fit$conf.int
```

```
[1] 27.723 28.236
attr(,"conf.level")
[1] 0.95
```

```
ttest.fit$p.value
```

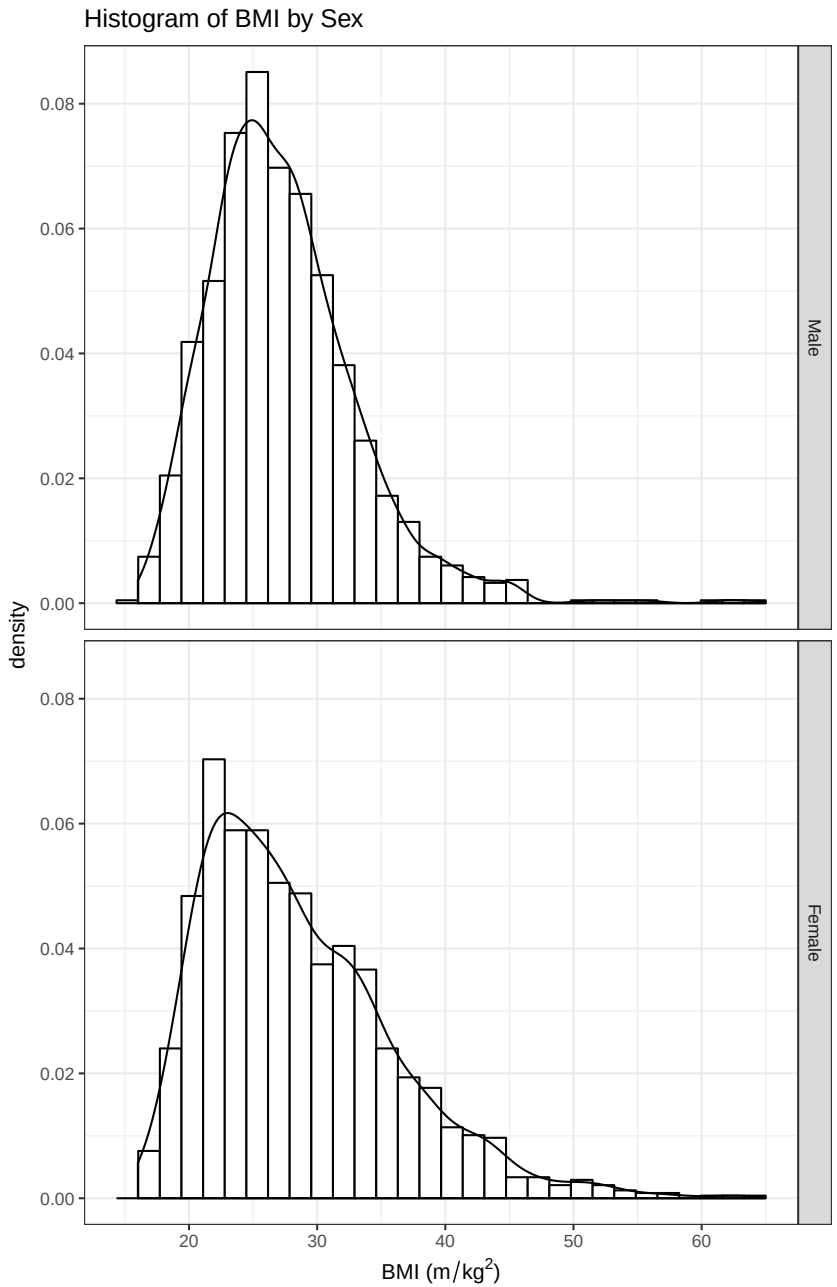
```
[1] 2.3446e-105
```

**Research question:** In American adults age 18-50, is the population mean BMI for males the same as for females?

# Visualization

```
# Create histogram of BMI by sex
ggplot(nhanes.adults, aes(x = bmi)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "white") +
  facet_grid(sex ~ .) +
  labs(title = "Histogram of BMI by Sex",
       x = expression(paste("BMI (", m/kg^2, ")"))) +
  geom_density() +
  theme_bw()
```

# Visualization



# Parameter estimation

- Statistical setup:
  - Let $X$ = BMI for males, $Y$ = BMI for females.
  - $X_1, ..., X_{n_m} \overset{iid}{\sim} (\mu_m, \sigma_m^2)$
  - $Y_1, ..., Y_{n_f} \overset{iid}{\sim} (\mu_f, \sigma_f^2)$
- Parameters/Estimators:
  - $\mu_\Delta = \mu_m - \mu_f$
  - $\widehat{\mu}_\Delta = \bar{X} - \bar{Y}$
  - 95% CI based on t-distribution $\Rightarrow$ 2 versions

# Parameter estimation

```
# Fit two-sample t-test by giving t.test two vectors
(ttest.fit <- t.test(nhanes.adults$bmi[nhanes.adults$sex == "Male"],
                     nhanes.adults$bmi[nhanes.adults$sex == "Female"]))



Welch Two Sample t-test

data:  nhanes.adults$bmi[nhanes.adults$sex == "Male"] and nhanes.adults$bmi[nha
t = -4.65, df = 2630, p-value = 3.6e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.70153 -0.69144
sample estimates:
mean of x mean of y
   27.351    28.548
```

# Parameter estimation

```r
# Fit two-sample t-test using formula notation (easier!)
(ttest.fit <- t.test(bmi ~ sex, data = nhanes.adults))
```

```
Welch Two Sample t-test

data:  bmi by sex
t = -4.65, df = 2630, p-value = 3.6e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.70153 -0.69144
sample estimates:
  mean in group Male mean in group Female
              27.351                28.548
```

# Hypothesis testing

Already saw results for two-sample t-test:

$$H_0 : \mu_\Delta = 0$$

$$H_A : \mu_\Delta \neq 0$$

$\Rightarrow$ Can also test whether $\mu_\Delta$ equals some non-zero value, but this is less common.

# In-class activity

(1) Find out whether we assumed equal variance.

(2) Decide whether we *should* assume equal variance.

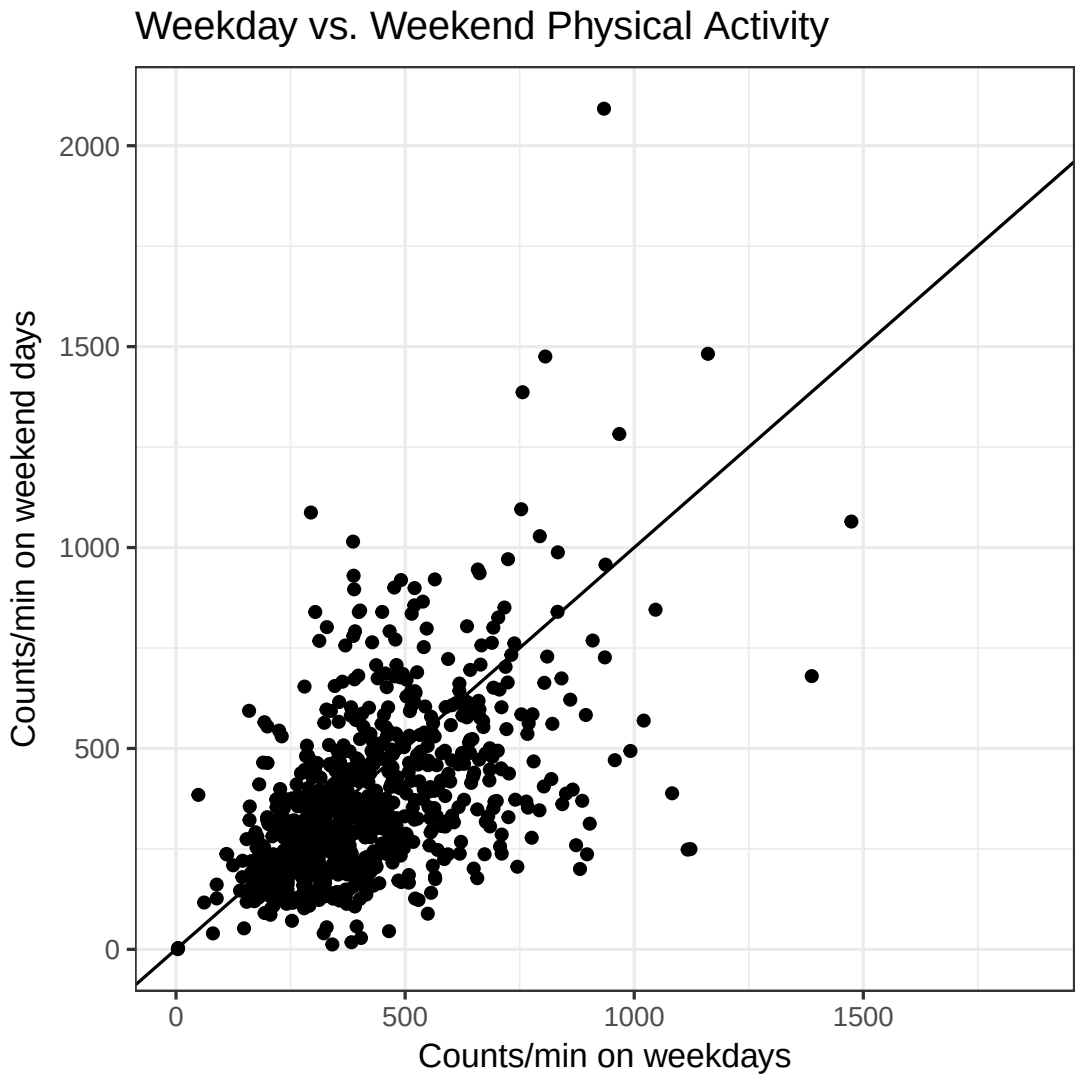
(3) Test $H_0 : \mu_\Delta = -1$, using appropriate test.

**Research question:** Are American adolescents age 13-17 more physically active on weekdays, or on weekend days?

# Visualization #1: Scatterplot

```
# Get subset of data for adolescents age 13-17
nhanes.adol <- subset(nhanes, age >= 13 & age <= 17)

# Plot weekend physical activity vs. weekday physical activity
ggplot(nhanes.adol, aes(x = wk_cpm, y = we_cpm)) +
  geom_point() +
  labs(title = "Weekday vs. Weekend Physical Activity",
       x = "Counts/min on weekdays",
       y = "Counts/min on weekend days") +
  geom_abline(intercept = 0, slope = 1) +
  theme_bw()
```

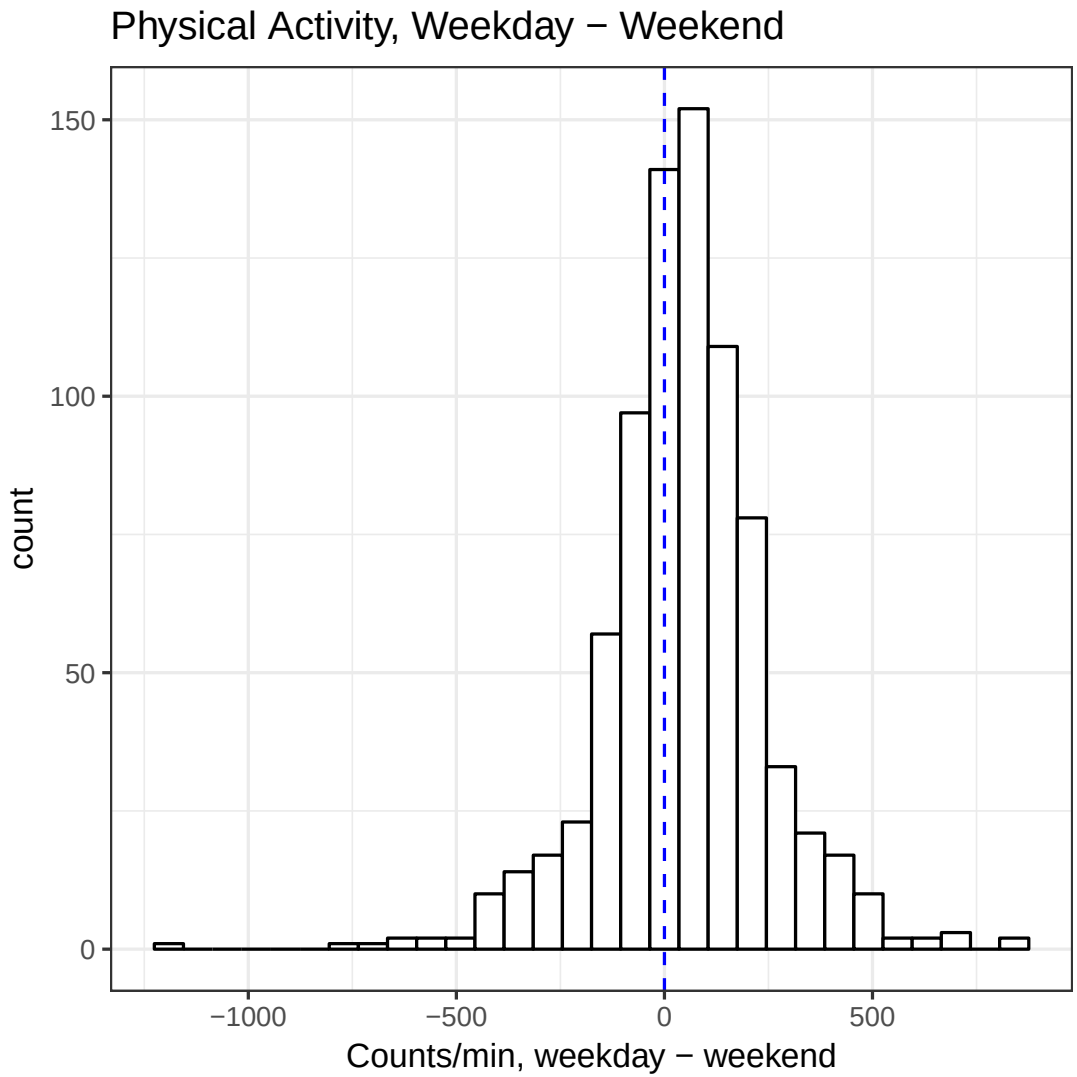
# Visualization #1: Scatterplot



# Visualization #2: Histogram

```r
# Calculate difference between weekday and weekend PA for each participant
nhanes.adol$cpm_diff <- nhanes.adol$wk_cpm - nhanes.adol$we_cpm

# Create histogram of differences
ggplot(nhanes.adol, aes(x = cpm_diff)) +
  geom_histogram(col = "black", fill = "white") +
  labs(title = "Physical Activity, Weekday - Weekend",
       x = "Counts/min, weekday - weekend") +
  geom_vline(aes(xintercept = 0),
             color = "blue", linetype = 2) +
  theme_bw()
```

# Visualization #2: Histogram



# Parameter estimation

- Statistical setup:
  - Let $X =$ Difference between average weekday PA and average weekend PA.
  - $X_1, ..., X_n \overset{iid}{\sim} (\mu, \sigma^2)$
- Estimators:

  $\hat{\mu} = \bar{X}$

  95% CI for $\mu$ : $\bar{X} \pm \frac{t_{(.975, n-1)} s}{\sqrt{n}}$

# Parameter estimation

```r
# Fit paired t-test by giving t.test two vectors
t.test(nhanes.adol$we_cpm, nhanes.adol$wk_cpm, paired = T)
```

```
Paired t-test

data:  nhanes.adol$we_cpm and nhanes.adol$wk_cpm
t = -6.7, df = 796, p-value = 3.9e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -60.469 -33.072
sample estimates:
mean of the differences
                 -46.77
```

# Parameter estimation

```r
# Fit paired t-test by giving t.test single vector of differences
t.test(nhanes.adol$cpm_diff)
```

```
One Sample t-test

data:  nhanes.adol$cpm_diff
t = 6.7, df = 796, p-value = 3.9e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 33.072 60.469
sample estimates:
mean of x
    46.77
```

# Hypothesis testing

Already saw results for paired t-test:

$$H_0 : \mu = 0$$

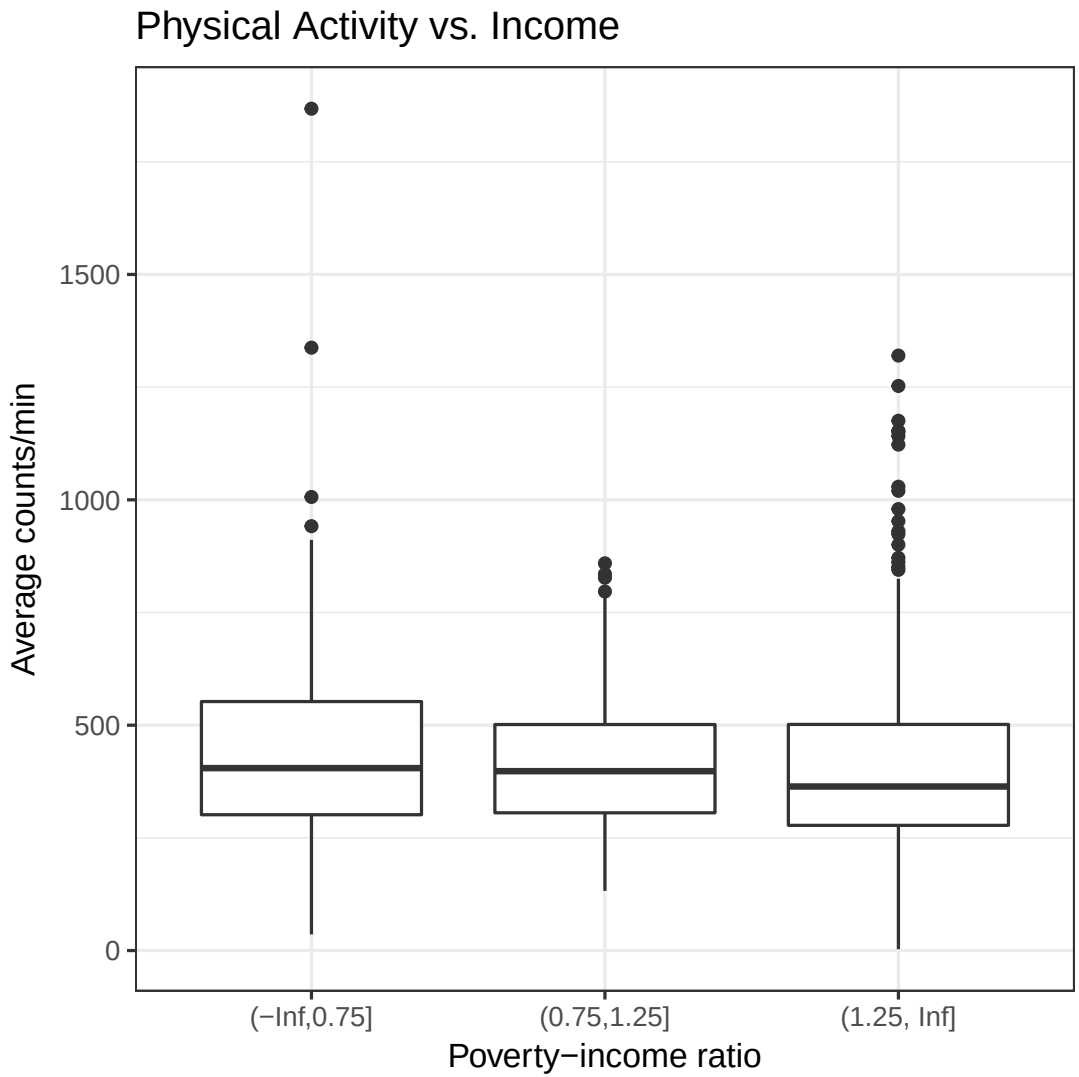
$$H_A : \mu \neq 0$$

**Research question:** Does physical activity differ by family income level (low, medium, high) in American adolescents?

# Visualization

```r
# Create three categories of poverty income ratio variable
nhanes.adol$pir.f <- cut(nhanes.adol$pir,
                         breaks = c(-Inf, 0.75, 1.25, Inf))

# Create boxplot of physical activity by PIR
ggplot(subset(nhanes.adol, !is.na(cpm) & !is.na(pir.f)),
       aes(x = pir.f, y = cpm)) +
  geom_boxplot() +
  labs(title = "Physical Activity vs. Income",
       y = "Average counts/min",
       x = "Poverty-income ratio") +
  theme_bw()
```

# Visualization



# Parameter estimation

- Statistical setup:
  - Let $X_{i,j}$ = Average physical activity for $j^{th}$ participant in $i^{th}$ PIR group, $i = 1, 2, 3$; $j = 1, ..., n_i$
  - Assume $X_{i,j} \overset{ind}{\sim} (\mu_i, \sigma^2)$, $i = 1, 2, 3$
- Estimators:
  - $\hat{\mu}_i = \bar{X}_i$
  - 95% CI for each $\mu_i$ : $\bar{X}_i \pm \frac{t_{(.975, n_i - 1)} s_i}{\sqrt{n_i}}$

# Parameter estimation

```
# Point estimates for mu's
tapply(nhanes.adol$cpm, nhanes.adol$pir.f,
       function(x) mean(x, na.rm = T))


(-Inf,0.75] (0.75,1.25] (1.25, Inf]
     438.33      414.47      409.44
```

# Parameter estimation

```
# Interval estimates for mu's
tapply(nhanes.adol$cpm, nhanes.adol$pir.f,
       function(x) t.test(x)$conf.int)


$`(-Inf,0.75]`
[1] 411.54 465.12
attr(,"conf.level")
[1] 0.95

$`(0.75,1.25]`
[1] 390.09 438.85
attr(,"conf.level")
[1] 0.95

$`(1.25, Inf]`
[1] 394.84 424.05
attr(,"conf.level")
[1] 0.95
```

# Hypothesis testing

Natural thing to test:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_A : \text{Not all } \mu's \text{ equal}$$

$\Rightarrow$ One-way ANOVA

# Hypothesis testing

```
# Fit ANOVA
anova.fit <- aov(cpm ~ pir.f, data = nhanes.adol)
summary(anova.fit)
```

```
              Df   Sum Sq Mean Sq F value Pr(>F)
pir.f          2   143785   71892    1.99   0.14
Residuals   1057 38259088   36196
231 observations deleted due to missingness
```

# Hypothesis testing

```
# Multiple comparisons
TukeyHSD(anova.fit)


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = cpm ~ pir.f, data = nhanes.adol)

$pir.f
                              diff     lwr     upr   p adj
(0.75,1.25]-(-Inf,0.75] -23.8621 -69.217 21.4927 0.43287
(1.25, Inf]-(-Inf,0.75] -28.8918 -63.014  5.2303 0.11582
(1.25, Inf]-(0.75,1.25]  -5.0297 -43.696 33.6368 0.94992
```

**Research question:** In older American males, Is there an association between race/ethnicity (4 levels) and obesity status (3 levels)?

# Generate variables

```
# Get subset of participants age 60+
nhanes.om <- subset(nhanes, sex == "Male" & age >= 60)

# Create 4-level factor version of race/ethnicity
nhanes.om$eth[nhanes.om$eth == 5] <- 4
nhanes.om$race.f <- factor(nhanes.om$eth, levels = 1: 4,
                           labels = c("Non-Hisp. White", "Non-Hisp. Black",
                                      "Mex. Amer.", "Other"))

# Create obesity variable
nhanes.om$obesity.f <- cut(nhanes.om$bmi,
                           breaks = c(-Inf, 25, 30, Inf), right = F,
                           labels = c("Normal", "Overweight", "Obese"))
```

# Visualization

```
# Contingency table with frequencies
(table.freq <- table(nhanes.om$race.f, nhanes.om$obesity.f))
```

```
                  Normal Overweight Obese
  Non-Hisp. White    124        209   131
  Non-Hisp. Black     34         43    31
  Mex. Amer.          36         83    45
  Other               16          7     5
```

# Visualization

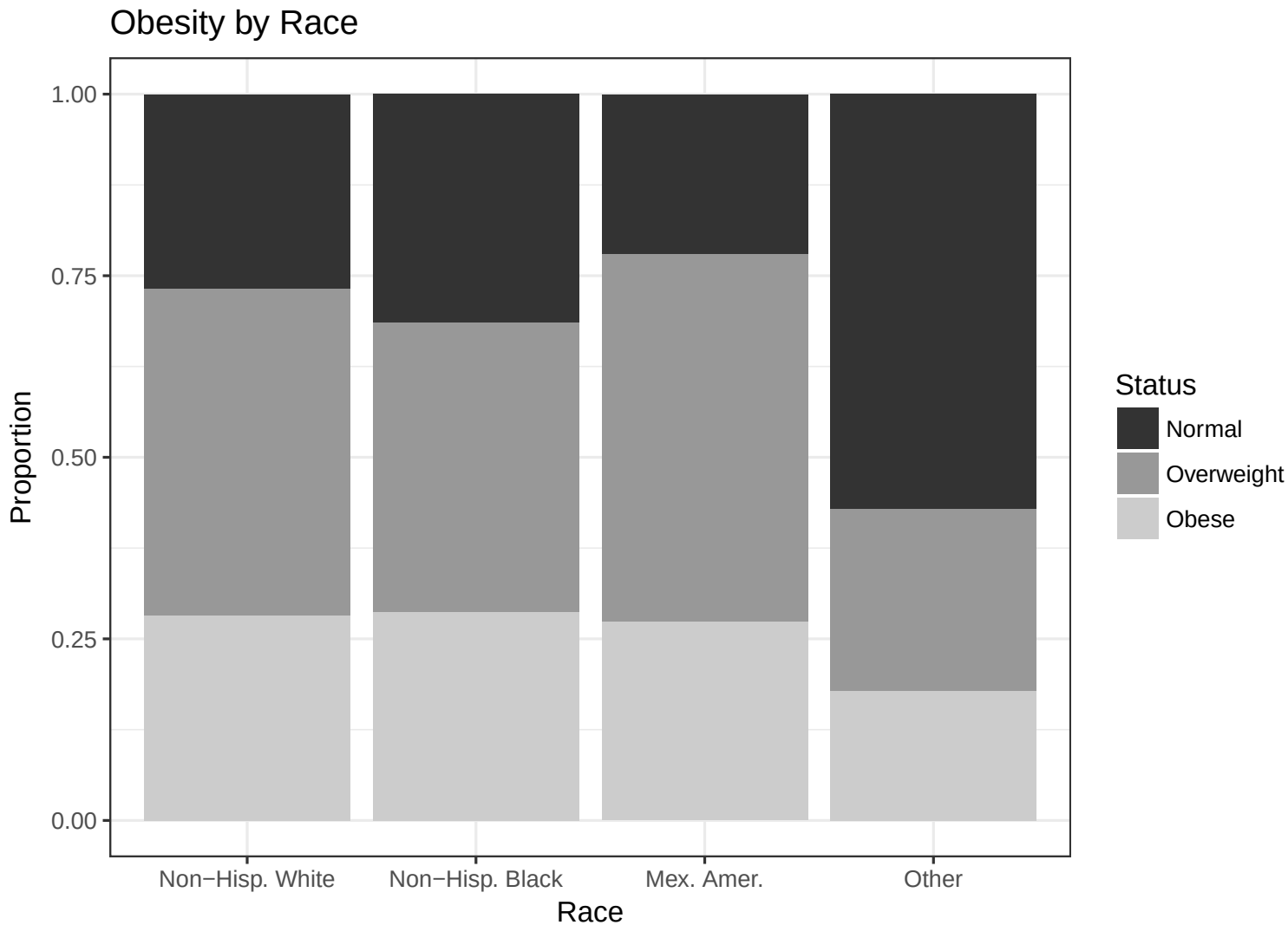
```
# Contingency table with row proportions
(table.rowprops <- prop.table(table.freq, margin = 1))
```

```
                  Normal Overweight   Obese
  Non-Hisp. White 0.26724    0.45043 0.28233
  Non-Hisp. Black 0.31481    0.39815 0.28704
  Mex. Amer.      0.21951    0.50610 0.27439
  Other           0.57143    0.25000 0.17857
```

# Visualization

```
# Create bar plot
ggplot(subset(nhanes.om, !is.na(race.f) & !is.na(obesity.f)),
       aes(x = race.f, fill = obesity.f)) +
  geom_bar(position = "fill") +
  labs(title = "Obesity by Race",
       y = "Proportion",
       x = "Race",
       fill = "Status") +
  theme_bw()
```

# Visualization



# Hypothesis testing

Typical test for two categorical variables:

$H_0$ : Race and obesity are not associated.

$H_A$ : Race and obesity are associated.

$\Rightarrow$ Chi-square test of association

# Hypothesis testing

```
# Chi-square test of association
chisq.test(nhanes.om$race.f, nhanes.om$obesity.f)
```

```
Pearson's Chi-squared test

data:  nhanes.om$race.f and nhanes.om$obesity.f
X-squared = 16.9, df = 6, p-value = 0.0098
```

# Hypothesis testing

```
# Chi-square test of association
chisq.test(nhanes.om$race.f, nhanes.om$obesity.f)
```

```
Pearson's Chi-squared test

data:  nhanes.om$race.f and nhanes.om$obesity.f
X-squared = 16.9, df = 6, p-value = 0.0098
```

$\Rightarrow$ P-value suggests race is significantly assoc. with obesity.

$\Rightarrow$ Chi-square test valid?

# Hypothesis testing

```
# Look at expected cell counts for each cell
(expected.counts <- matrix(rowSums(table.freq), ncol = 1) %*%
   colSums(table.freq) / sum(table.freq))
```

```
         [,1]    [,2]     [,3]
[1,] 127.5393 207.707 128.7539
[2,]  29.6859  48.346  29.9686
[3,]  45.0785  73.414  45.5079
[4,]   7.6963  12.534   7.7696
```

# In-class activity

(1) Figure out what small-sample test we could use here.

(2) Figure out what R function does it.

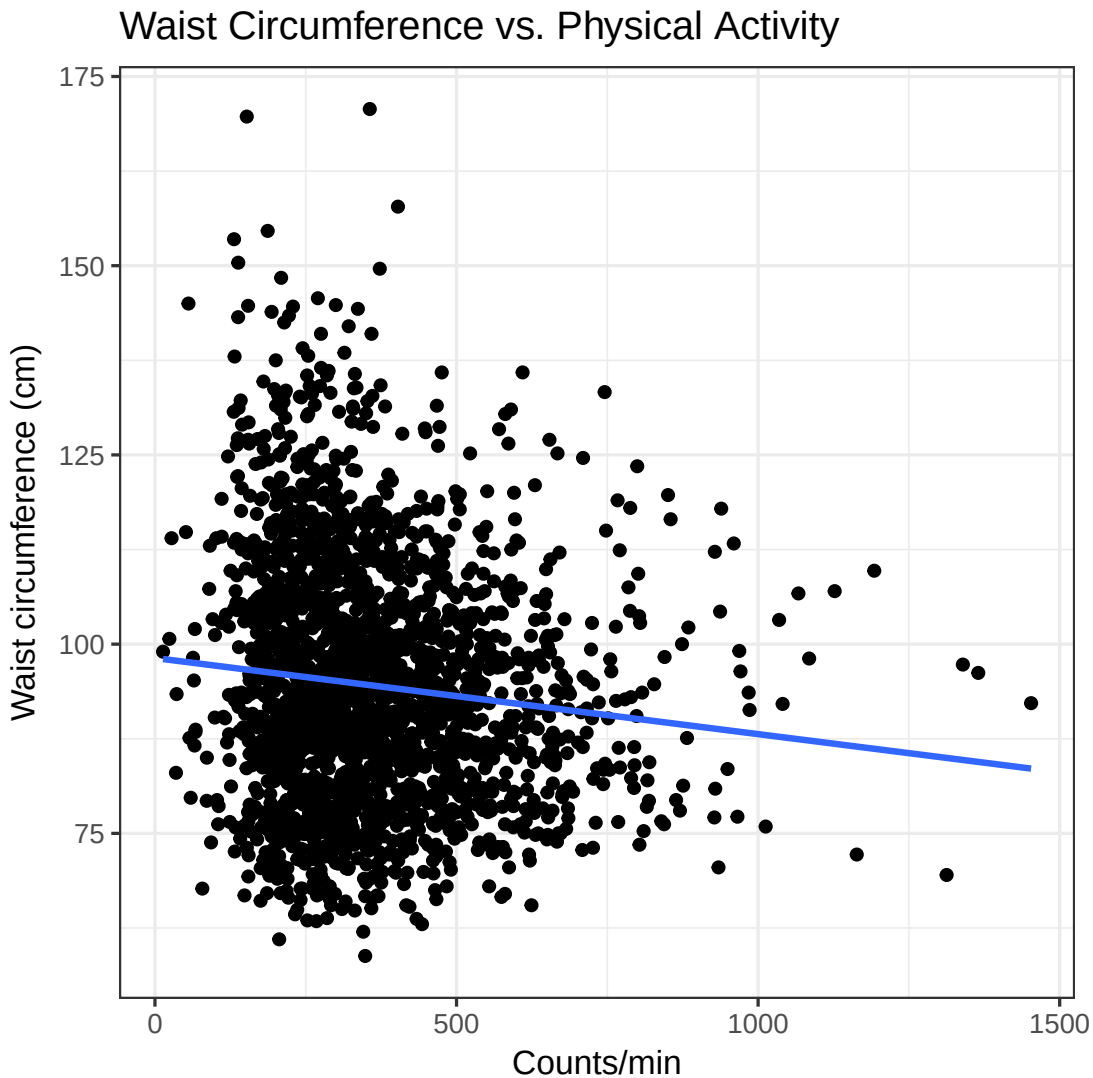
(3) Perform test $\Rightarrow$ same conclusion as Chi-square?

**Research question:** In American adults age 18-50, is there a correlation between physical activity and waist circumference?

# Visualization

```
# Scatterplot of waist circumference vs. physical activity
ggplot(nhanes.adults, aes(cpm, waist)) +
  geom_point() +
  labs(title = "Waist Circumference vs. Physical Activity",
       x = "Counts/min",
       y = "Waist circumference (cm)") +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw()
```

# Visualization



# Correlation analysis

- Statistical setup:
  - Let $X$ = physical activity, $Y$ = waist circ.
  - Assume $(X_1, Y_1), ..., (X_n, Y_n) \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Parameter of interest:

$$\rho_{xy} = \text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E(XY)-E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

- Hypothesis test:

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

# Correlation analysis

```
# Calculate Pearson correlation coefficient
cor.test(nhanes.adults$cpm, nhanes.adults$waist)
```

```
Pearson's product-moment correlation

data:  nhanes.adults$cpm and nhanes.adults$waist
t = -5.06, df = 2270, p-value = 4.6e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.146034 -0.064721
sample estimates:
     cor
-0.10555
```

# Correlation analysis

```r
# Calculate Spearman correlation coefficient
cor.test(nhanes.adults$cpm, nhanes.adults$waist, method = "spearman")
```

```
Spearman's rank correlation rho

data:  nhanes.adults$cpm and nhanes.adults$waist
S = 2.16e+09, p-value = 9.1e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
-0.10275
```

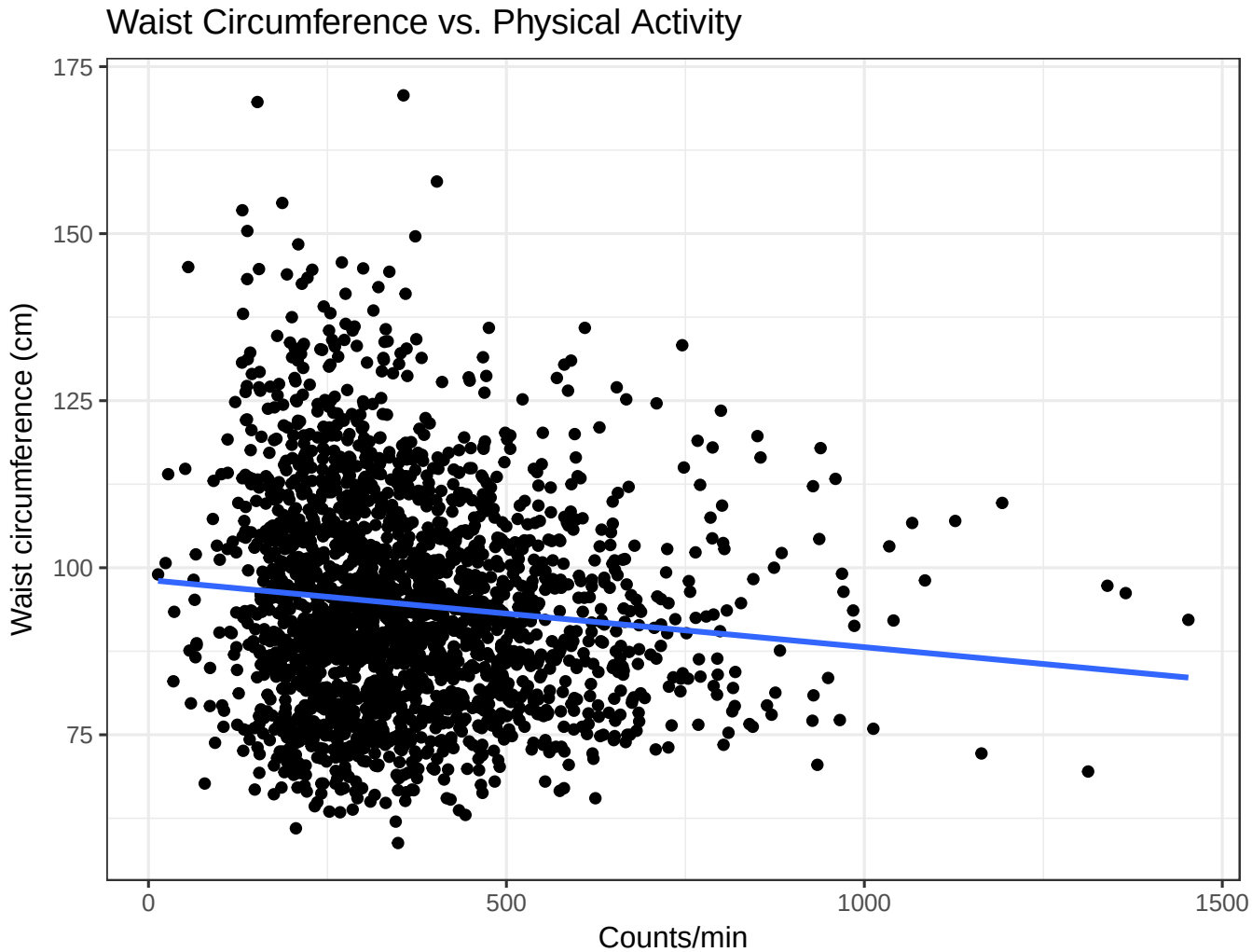
# Regression analysis

- Statistical setup:
  - Assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$
- Parameters/estimation:
  - $\beta_0 =$ intercept, $\beta_1 =$ slope
  - OLS: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$
- Hypothesis test of primary interest:

  $H_0 : \beta_1 = 0$

  $H_A : \beta_1 \neq 0$

# Intercept and slope



# Regression analysis

```r
# Fit linear regression for waist circumference vs. physical activity
linear.fit <- lm(waist ~ cpm, data = nhanes.adults)
summary(linear.fit)
```

# Regression analysis

```
Call:
lm(formula = waist ~ cpm, data = nhanes.adults)

Residuals:
   Min     1Q Median     3Q    Max
-35.86 -11.74  -1.65   9.72  76.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.16700    0.79261  123.85  < 2e-16 ***
cpm         -0.01006    0.00199   -5.06  4.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.9 on 2271 degrees of freedom
  (440 observations deleted due to missingness)
Multiple R-squared:  0.0111, Adjusted R-squared:  0.0107
F-statistic: 25.6 on 1 and 2271 DF,  p-value: 4.57e-07
```

# Regression analysis

```r
# Divide CPM by 100 to make slope easier to interpret
nhanes.adults$cpm_100 <- nhanes.adults$cpm / 100
linear.fit <- lm(waist ~ cpm_100, data = nhanes.adults)
summary(linear.fit)
```

# Regression analysis

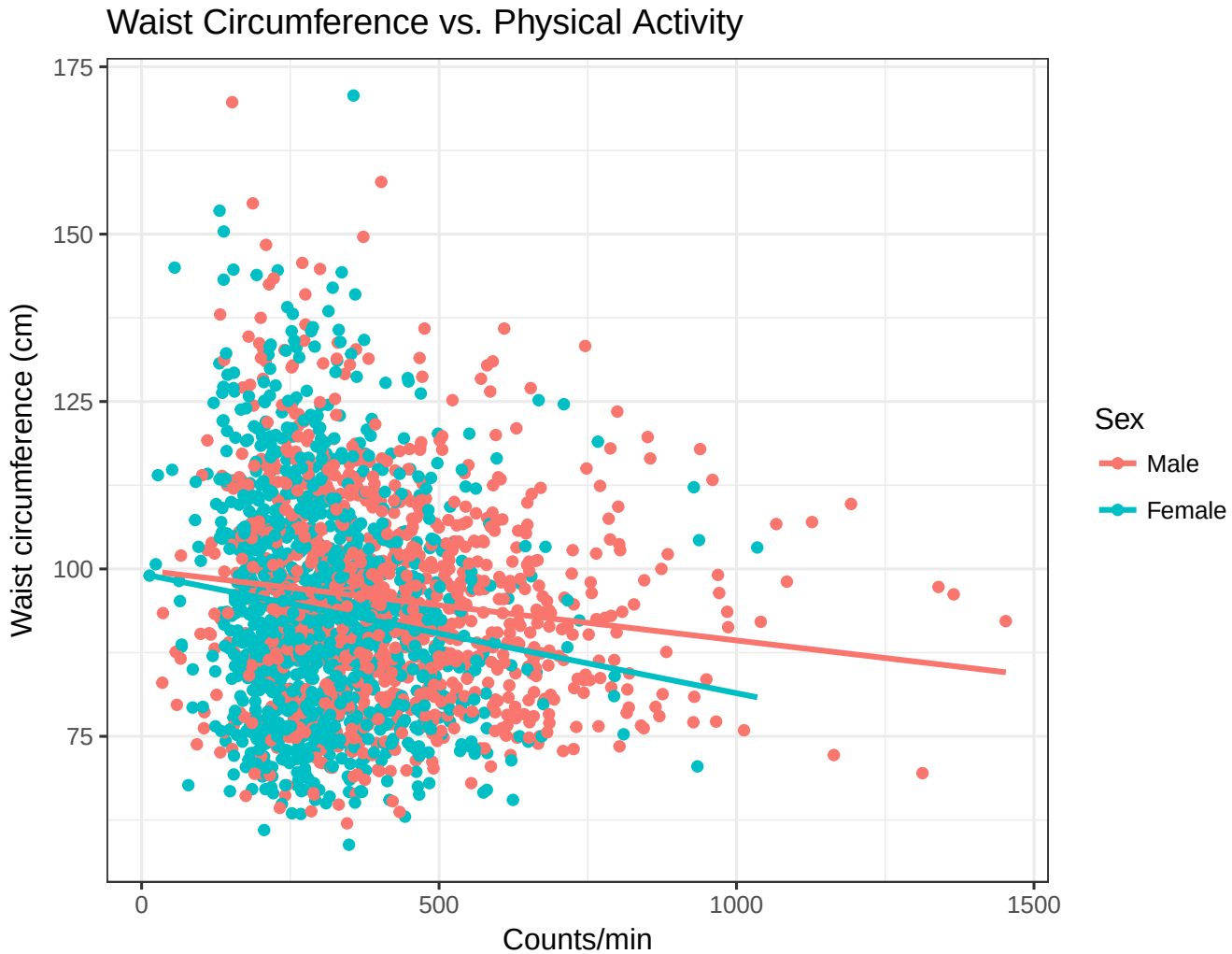
```
Call:
lm(formula = waist ~ cpm_100, data = nhanes.adults)

Residuals:
   Min     1Q Median     3Q    Max
-35.86 -11.74  -1.65   9.72  76.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   98.167      0.793  123.85  < 2e-16 ***
cpm_100       -1.006      0.199   -5.06  4.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.9 on 2271 degrees of freedom
  (440 observations deleted due to missingness)
Multiple R-squared:  0.0111, Adjusted R-squared:  0.0107
F-statistic: 25.6 on 1 and 2271 DF,  p-value: 4.57e-07
```

**Research question:** Does the relationship between physical activity and waist circumference differ by sex?

# Visualization

```
# Scatterplot of waist circumference vs. physical activity
ggplot(nhanes.adults, aes(cpm, waist, color = sex)) +
  geom_point() +
  labs(title = "Waist Circumference vs. Physical Activity",
       x = "Counts/min",
       y = "Waist circumference (cm)") +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw()
```

# Visualization



# Regression analysis

- Statistical setup:
  - Let $X$ = physical activity, $Y$ = waist circumference, and $M = 1$ if male, 0 if female.
  - Assume
    $Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i M_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$
- Hypothesis test of primary interest:

  $H_0 : \beta_3 = 0$

  $H_A : \beta_3 \neq 0$
- What does it mean if $\beta_3 = 0$?

# Regression analysis

```r
# Fit model with interaction term
nhanes.adults$male <- ifelse(nhanes.adults$sex == "Male", 1, 0)
linear.fit <- lm(waist ~ cpm + male + cpm * male, data = nhanes.adults)
summary(linear.fit)
```

# Regression analysis

```
Call:
lm(formula = waist ~ cpm + male + cpm * male, data = nhanes.adults)

Residuals:
   Min     1Q Median     3Q    Max
-34.59 -11.55  -1.77   9.45  77.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 99.26133    1.22663   80.92   <2e-16 ***
cpm         -0.01782    0.00364   -4.90    1e-06 ***
male         0.56108    1.68716    0.33    0.739
cpm:male     0.00732    0.00443    1.65    0.099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.8 on 2269 degrees of freedom
  (440 observations deleted due to missingness)
Multiple R-squared:  0.0208, Adjusted R-squared:  0.0195
F-statistic: 16.1 on 3 and 2269 DF,  p-value: 2.4e-10
```

# Regression analysis

**Question:** What if we drop the interaction term? It was not significant after all.

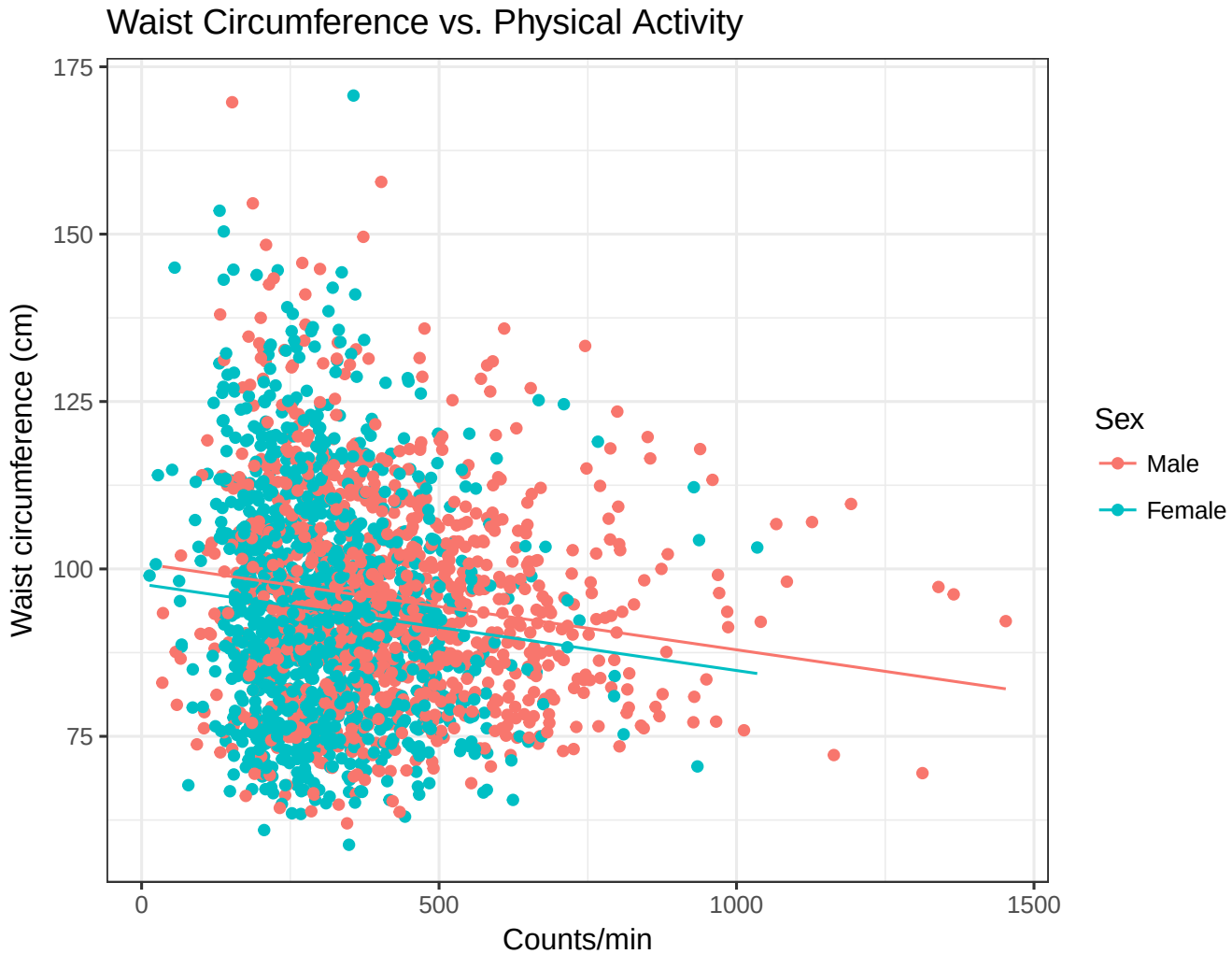
# Regression analysis

- Previously, with interaction term:
  - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i M_i + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$
- Now:
  - $Y_i = \beta_0^* + \beta_1^* X_i + \beta_2^* M_i + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} (0, \sigma^{*2})$
- Parameters:
  - Now, what is the slope for males? Females?
  - Are the regression lines the same?

# Visualization

```
# Scatterplot with regression line that does not include interaction
fit <- lm(waist ~ cpm + as.factor(male), data = nhanes.adults)
nhanes.adults$pred_waist <- predict(fit, newdata = nhanes.adults)

ggplot(nhanes.adults, aes(cpm, waist, color = sex)) +
  geom_point() +
  labs(title = "Waist Circumference vs. Physical Activity",
       x = "Counts/min",
       y = "Waist circumference (cm)",
       color = "Sex") +
  geom_line(aes(y = pred_waist)) +
  theme_bw()
```

# Visualization



# Regression analysis

```
# Fit model without interaction term
linear.fit <- lm(waist ~ cpm + male, data = nhanes.adults)
summary(linear.fit)
```

# Regression analysis

```
Call:
lm(formula = waist ~ cpm + male, data = nhanes.adults)

Residuals:
   Min     1Q Median     3Q    Max
-34.43 -11.63  -1.82   9.47  77.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.72059    0.79575  122.80  < 2e-16 ***
cpm         -0.01289    0.00208   -6.20  6.8e-10 ***
male         3.09604    0.69763    4.44  9.5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.8 on 2270 degrees of freedom
  (440 observations deleted due to missingness)
Multiple R-squared:  0.0196, Adjusted R-squared:  0.0188
F-statistic: 22.7 on 2 and 2270 DF,  p-value: 1.66e-10
```

# Final thoughts

- The internet exists. No need to memorize!
- Function help files are...helpful.
- In most cases, a graph is (at least) as good as a test.
- Make graph $\Rightarrow$ eyeball association $\Rightarrow$ perform test.

# Some other topics

- Survival analysis (**survival** package)
- Longitudinal analysis (**nlme**, **lme4**, **gee** packages)
- Complex survey analysis (**survey** package)

# Lab

You may or may not be familiar with logistic regression. Logistic regression is what you use to test whether one or more variables are associated with a binary outcome variable.

For logistic regression with a binary outcome $Y$ and two predictors $X_1$ and $X_2$, we assume the following model:

$$\log[\frac{P(Y_i=1)}{1-P(Y_i=1)}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

In other words, we assume there is a linear relationship between each predictor and the log-odds of $Y$.

If $\beta_1 = 0$, then $X_1$ is not associated with $Y$. If $\beta_1$ is positive, then people with higher values for $X_1$ are more likely to experience the outcome than people with lower values of $X_1$; and vice versa if $\beta_1$ is negative.

In this lab, you will learn how to use the *glm* function to fit a logistic regression model in R. It works like *lm*, but can handle various regression models, not just linear regression.

Follow these steps to test whether sex and waist circumference are associated with odds of meeting the US physical activity guidelines for adults ($\geq$ 150 minutes of exercise per week).

1. Run code from slide 6 to download **nhanes** data frame.
2. Subset data for participants age 18-65.
3. Create variable that is 1 if **guideline_min** $>$ 21.4, and 0 otherwise.
4. Look at help file for *glm* and see what input you have to specify to get R to do logistic regression.
5. Fit and interpret logistic regression for meeting guidelines vs. sex and waist circumference.
6. Re-fit model with sex-by-waist circumference interaction.

   Does relationship between waist circumference and odds of meeting the guidelines differ by sex?

   If so, is the association stronger in males or females?